Method to cut a tree into subtrees to define subfamilies

This software is designed to divide a family into subfamilies. It is the basic idea behind the Kerf code (used within SATCHMO-JS to jump-start the SATCHMO algorithm).

In implementing this algorithm, it is okay to use any open-source code, but just cite the code source.

Aim: Given an input MSA and corresponding tree, and a percent identity cutoff X, produce a set of MSAs and subtrees such that no pair of sequences in any individual sub-MSA has <X% pairwise sequence identity.

Input: MSA, tree, X (real value (0 < X < 100)).

Output: (1) A CSV file with one row (line) for each sequence in the input MSA, with the subfamily number (auto-increment), sequence identifier. (2) a set of MSAs, one for each subfamily identified. Each subfamily MSA should be drawn from the input MSA (i.e., take the rows from the input MSA and print them to a separate file). Confirm that the MSAs you produce can be viewed using a standard MSA viewer such as Jalview.

Sample data can be downloaded from: http://phylogenomics.berkeley.edu/book/book_info.php?book=bpg087857&short_name=APAF

Process: Read in the MSA and tree. Starting at the leaves, compute the minimum pairwise percent identity over all pairs of sequences within each subtree.
Store the value for each subtree node. If a subtree node is reached where that minimum pairwise identity drops below the threshold X, break off the child subtrees as separate subtrees. Repeat until all leaves are included in one of the output subtree/MSAs.

Note that percent identity is defined as: #exact matches of amino acids / #positions where at least one sequence aligns. (i.e., do not consider positions where both sequences have gaps).

Example toy MSA:

```
1: MSTPP----W
2. -TTPPPPP-W

The number of identities is 4.
The number of positions where both sequences align is 8.
The two sequences have 50% identity.
```